A Defense of Temperate Epistemic Transparency

Eleonora Cresto – CONICET (Argentina)
eleonora.cresto@gmail.com

May 2011

1. Introduction

Epistemic logic is typically concerned with the formalization of crucial aspects of the concept of knowledge. Once basic features of our pre-theoretical understanding of knowledge are captured by suitable axioms, we can go on to explore the formal properties of the system, which sometimes lead us to discover important consequences that might not be so evident at first sight. As with many theories (both formal and informal), we look for a reflective equilibrium between our prior intuitions and the claims that turn out to be valid in the theoretical framework. And, as with many other modal logics, we can discuss which particular system best expresses the features we have in mind.

In particular, we can wonder whether an adequate account of knowledge is such that, from the fact that an agent knows a given proposition *p*, we can infer that she also knows that she knows that *p*, i.e., we can wonder whether we have the right to infer that she has *second-order* knowledge. This is precisely the idea encoded in the so-called ***KK principle***, or $Kp \rightarrow KKp$. It is also referred to as a *principle of positive introspection*, *principle of knowledge-reflexivity*, *epistemic transparency*, *self-knowledge* or *luminosity*; in what follows I will refer to ***KK*** by any of these names indistinctly. In this paper I will attempt a defense of a temperate version of ***KK*** as a normative, rather than descriptive, epistemic principle.

Epistemic transparency can well be taken to be implicit in traditional theories of knowledge, from Plato to the first half of the Twentieth Century. Once epistemic logics appeared in the epistemological scene (starting with Hintikka's seminal book *Knowledge and Belief*, in 1962), the ***KK*** principle became explicitly stated in the formal literature. Then epistemic externalism broke in, and things began to change. Externalism tells us, precisely, that justification is not "internal" to our consciousness, but is a matter of placing ourselves in the right relationship with the environment, regardless of

whether we are aware of this fact. Here's a delicate point, however, because to say that knowledge does not require any particular state of awareness regarding justification does not necessarily indicate that agents will typically know without knowing that they know, insofar as second-order knowledge could *also* be defined in an externalistic fashion. Thus, assuming an agent formed the corresponding second-order belief, such a belief could be justified by reasons to which, once again, she has no access.

Of course, externalism also allows for the idea that knowledge does not demand the corresponding second-order belief to begin with; clearly, in this case positive introspection fails. In any case, there are well-discussed, explicit rejections of transparency that proceed through more sophisticated paths. Many such arguments involve an attempt to escape from first- order skepticism: by rejecting ***KK*** we prevent second-order skeptical claims to infest the first level through *Modus Tollens*.

In recent years we have witnessed the development of new arguments against ***KK*.** Timothy Williamson, in particular, articulates two fundamental lines of attack. On the one hand he suggests, mostly through carefully chosen examples, that no mental state is luminous, not even the most "obvious" ones – hence surely knowledge is not luminous either.[1] On the other hand, Williamson contends that knowledge requires a margin of error: if I am entitled to say that I know that the tree in front of me is not 0.5 meters, then the tree should not be 0.51 meters either. This requirement can be seen as a particular elaboration, for perceptual knowledge, of the so-called *safety* condition. The safety condition tells us that what the agent believes should be true not only in the actual world, but also in close worlds the agent cannot discriminate from the actual one. Williamson then shows, by means of a soritic argument, that if we accept ***KK***, the margin of error principle leads us to an inconsistency (cf. Williamson 2000, chapter 5).

I believe Williamson's considerations on margin of error principles reveal us something important about the structure of first-order knowledge. However, there is still room to propose a limited defense of ***KK***.[2] I will argue in favor of a moderate version of ***KK*** indirectly, by showing that it is forced upon us as a result of a number of independent decisions on the structure of second-order probabilities and the way

[1] Thus, for instance, a hypochondriac could falsely think she is in pain (cf. Williamson 1995, p. 535; cf. also Sosa 2007, chapter 2.) For an interesting discussion on this point see Leitgeb (2002).

[2] An important antecedent of what I intend to do here can be found in Egré and Dokic (2008), and Egré (2009). I will come back to their proposal in Section 8 of this paper.

probability and knowledge attributions cohere with each other; it should be noted that my argument will not be tied to an internalist project about justification.

The article is organized as follows. It comprises two main parts, quite distinct from one another. In the first part (section 2) I consider briefly why we should focus on epistemic transparency in the first place. Notice that my answer to the question of why we care about epistemic transparency is not the central argument of the paper: it is just meant to motivate the search for the results this work seeks to obtain. The second part (sections 3 to 8) concerns the (indirect) argument for epistemic transparency properly speaking. My starting point will be Williamson's formal proposal in "Improbable Knowing" (section 3); then I suggest an alternative setting that can be deemed philosophically satisfactory (sections 4 and 5); finally, I devote some space to discuss the philosophical consequences of the model (sections 6 and 7), as well as its connection with recent work on the topic (section 8).

First Part: Theoretical Motivations

2. Why Should We Care? Rationality, Responsibility and Reflection

Why is positive introspection important? One can be tempted to contend here that epistemic transparency is simply a feature of idealized reasoners, such that an epistemic system should validate transparency if and only if the system explicitly deals with ideally rational agents. According to this line of thought, either attempting to defend a principle of epistemic transparency is idle or impossible (because real agents do not satisfy it). This strikes me as a false dichotomy. I would like to suggest that introspective principles do much more (and, in a different sense, much less) than depict an ideal reasoner.

The assumption of ideal rationality deserves a careful discussion, of which here I can offer no more than an outline. We can put into question, for instance, whether epistemic models should always presuppose logical omniscience, where "logical omniscience" usually refers to consistency and deductive closure; sometimes the term is

also meant to include probabilistic coherence, and perhaps even expected utility maximization. At one end of the spectrum we find philosophers who defend at least part of such requirements without reservations; we can say, for instance, that logical omniscience embodies a regulative ideal that we clearly wish to fulfill. Thus, as soon as we notice that a particular proposition follows deductively from some of our beliefs, we typically feel compelled to believe it as well – or to revise our background. Moreover, as deduction transmits truth, recognizing that we know the premises usually forces us to recognize that we also know the conclusion. This account seems particularly sensible if we understand beliefs as *commitments*, as some authors have proposed. At the opposite end of the spectrum we find defenders of so-called *bounded rationality*. Most supporters of bounded rationality rely on some version of the "ought-implies-can" maxim, and hence argue that we should not ask for what we are, as a matter of fact, unable to attain.

In any case, we could well distinguish the ideal of epistemic rationality that gets captured by logical omniscience from the type of idealization that springs from the acceptance of (some version of) epistemic transparency. Even if logical omniscience were deemed to be an appropriate demand (say, for reasons that ultimately go back to our pre-theoretical understanding of rationality), it does not follow from here that transparency demands will be found just as appropriate. Consider, for instance, the many formal proposals that rely on a Kripkean-based semantics – which forces agents to be logically omniscient – while at the same time rejecting the validity of ***KK***. I take it that such proposals are implicitly committed to the modeling of ideally rational agents; however, in such frames, knowing that one knows remains conceptually closer to empirical rather than to *a priori* knowledge, and hence it can be subjected to empirical doubts. Incidentally, recall that in general we do not ask agents to satisfy empirical omniscience to be rational.[3]

I would like to seek a middle ground between those who take transparency to be part of our ideal of a rational epistemic agent,[4] and those who refuse to do so. The position favored in this paper is then intermediate: on the one hand, I agree that transparency is not an ideal of rationality in the same sense that consistency or

[3] For a discussion of this point *cf.* Christensen (2004), chapter 6.

[4] For an explicit defense of the view that transparency is part of our ideal of a rational agent cf. Levi (1997), among others. Assuming we can come up with a suitable linguistic representation of our epistemic states, Levi argues that, were we to enrich the representation language so as to be able to talk about our own knowledge states, the rational system to pick would validate the ***KK*** principle.

deductive closure are; however, this is not to say that it is not an important ideal on its own right, albeit of a different type: it is an ideal of *epistemic responsibility*.[5]

How does the connection between transparency and responsibility go, exactly? Some authors have contended that the mere fact that we have certain beliefs already entitles us to talk about epistemic responsibility, in some deflated sense.[6] But, as I see it, responsibility is hardly an all-or-nothing affair. It might be correct to say that, at times, all we require from agents in order to credit them with responsibility is to have certain doxastic attitudes. However, there are uses of the concept of epistemic responsibility tht are not so minimal. In particular, there is a clearly identifiable sense of the term according to which we are reluctant to say that an agent is fully epistemically responsible for her belief that *p* unless she is very much aware of her having *p*, and perseveres on her belief that *p* on reflection. Part of the explanation for our reluctance is that agents who fully embrace their first-order mental states are perceived as more in control of themselves, and as having a more sophisticated epistemic life. This richer sense of responsibility is clearly a desideratum. To put it in a slogan: "make sure you own your own beliefs [and desires]". Transparency, under some suitable formulation to be discussed in the sections to follow, is then a requirement of *fully idealized* responsible agents: if *S* is a fully idealized responsible agent and *S* knows that *p*, then *S* has duly reflected on her (first-order) epistemic states and has found it to be the case that she knows that *p*.

Many philosophers exploited the link between agenthood and transparency in the past. A possible way to go would be to contend that having knowledge of our own intentional states is *constitutive* of the very idea of intentional state.[7] I don't think we

[5] Thus, according to the view I favor, epistemic responsibility and epistemic rationality can very well come apart – and they often do. Most authors simply assume that they amount to more or less the same thing (cf. for example Owens's description of what he dubs the "juridical theory of responsibility", in Owens (2000)).

[6] Cf. Engel (mn), Hieroyimi (2005), or Owens (2000), among others. According to this tradition, discussions about epistemic responsibility lead us to the problem of epistemic voluntarism. Thus we could be tempted to reason as follows: beliefs are voluntary only in a much deflated sense, but nonetheless we can be subjected to criticism for having a particular doxastic corpus rather than other; hence the necessary conditions for epistemic responsibility cannot be too stringent. In Owens (2000) we find an interesting attempt to preserve a deflated sense of responsibility even under the assumption that agents have no freedom whatsoever regarding their own doxastic states.

[7] Cf. for example Bilgrami (1988). Bilgrami argues that we should follow Strawson in thinking of freedom as not purely metaphysical, but normative. Freedom is then defined by the "reactive attitudes" (blame, criticism, resentment) we find in ourselves and in others; moreover, according to Bilgrami it is also defined by the normative reactions we can *justify* with our values. He then contends that we cannot justify our criticism of an agent's beliefs or desires unless we assume the agent to have self-knowledge of

need to commit ourselves to anything this strong, though. We can accept that full-fledged responsibility requires being in the right sort of reflective state, without pronouncing ourselves as to whether this fact is actually constitutive of intentional attitudes.[8] In any case, we have to be careful concerning what sense of reflective state is at stake. We might wonder, for instance, whether the reflective stance that I identified as necessary for full-fledged epistemic responsibility is not just a side effect of the demand for epistemic justification, particularly as understood by internalist epistemology. I take it that the answer is "no". According to the sense of reflection that gets vindicated by internalist epistemology, epistemic responsibility is a trait which, when possessed by the agent, prevents her from believing without due respect for the empirical evidence. According to the sense I am interested in, by contrast, epistemic responsibility is a trait that forces an agent to fully embrace her first-order beliefs: the agent can then be said to "ratify" them. Sometimes the two phenomena go together, but we can also have the second without the first.

To be successful, this picture has yet to tell us how the reflective stance involved in ratification (in the sense just explained) can give us second-order *knowledge*, in addition to second-order belief. I cannot elaborate on this idea here, due to space limitations, but let me just give a hint as to how the argument would go. The crucial step is to recognize that justification is not always perceived as necessary for knowledge attribution. If we pay due attention to the linguistic evidence we will notice that, under certain circumstances, agents do not care about justification at all, and yet they are naturally inclined to talk about knowledge. This is particularly noticeable in cases in which the attributor already believes that *p* is true; in such cases, finding out that the agent also believes that *p* does not provide information about *p* (as far as the attributor is concerned) but about the agent. In many such circumstances, then, we can observe that to attribute knowledge that *p* to a subject *is* just to attribute true belief that *p*. A possible explanation for this fact is that justification becomes relevant when – but only when –

---

her own intentional states. Thus, self-knowledge is a necessary condition for responsibility, and the following conditional holds: "To the extent that an intentional state is in the region of responsibility, i.e., to the extent that an intentional state is the rational cause of an action which is the object of justifiable reactive attitudes, or to the extent that an intentional state is itself the object of a justifiable reactive attitude, then that intentional state is known to its possessor." (p. 8). Exceptions to self-knowledge are precisely signs of the inapplicability of the normative conditions specified in the antecedent.

[8] Cf. Chapter 1 of Foley (2001) for an alternative account on epistemic responsibility that also puts reflection at the center stage.

we enter a very particular reflective stance, to wit, *when we examine the relevant beliefs under the light of a possible revision*.[9] But the ability to have or lack justification does not preexist: the conceptual space for justification is *created* by placing ourselves in what we might call "a deliberation mood". Reflecting on our beliefs, or on the beliefs of a third party agent, need not involve a deliberation mood. Thus not every reflective stance is a justification stance. In light of this, it is not generally true that full-fledged epistemic responsibility demands justification (although sometimes it certainly does). Sometimes the quest for justification just doesn't arise, but the agent could still be said to have knowledge, and second-order knowledge, of the relevant propositions – and hence qualify as fully responsible.[10]

In what follows I will seek to show that we have formal reasons to argue in favor of some version of positive introspection, and hence in favor of a model that captures the concept of an ideally responsible agent. This fact, I take it, should reinforce the claim that epistemic transparency is a desideratum we should not be too ready to dismiss.

Second Part: A Formal Argument in Favor of Epistemic Transparency

3. Williamson on Improbable Knowing

Let me start by recalling some of the axioms that are often discussed when we formulate an epistemic logic. It is usually accepted, for instance, that all tautologies from propositional logic should be valid in our system. Consider next axiom **K**:

K $\quad K(\varphi \rightarrow \psi) \rightarrow (K\varphi \rightarrow K\psi)$

[9] These suggestions are best read as particular re-elaborations of what some authors have dubbed "the Belief-Doubt model", of Peircean roots, according to which we should not devote our energies to justify prior beliefs, but to justify belief *changes* (cf. chapter 1 of Levi (1997), among others).

[10] Just to clarify, the claim is not that, if *S* knows that *p*, then *S* will always know that she knows that *p*, regardless of whether her second order belief is or is not justified. Rather, the claim is: if *S* is ideally responsible and *S* knows that *p*, then (*S* believes that she knows that *p*, and (either *S* is justified in believing that *p*, or *S's* belief is not in the conceptual space required for justification to be meaningful in the first place)).

K amounts to saying that, if an agent knows both an implication and its antecedent, then she also knows the consequent. **K** is valid in any "normal system" (in Kripke's sense); those who argue against deductive closure for knowledge, such as Robert Nozick (1981), are bound to reject it. Consider also:

T $K\varphi \rightarrow \varphi$

T says that if someone knows that φ, then φ is the case (we cannot know false things); **T** seems reasonable if we think, as most people do, that knowledge is "factive", i.e., it involves truth. Finally, we may also wonder about the validity of a number of introspective principles, such as ***KK***, as well as the so-called *principle of wisdom*, or *principle of negative introspection*:

Positive introspection (*KK*) $K\varphi \rightarrow KK\varphi$

Negative introspection $\sim K\varphi \rightarrow K\sim K\varphi$

Most authors have found negative introspection to be even more contentious than ***KK*;** I will not deal with it any further in this paper.

On the other hand, in order to provide a semantic machinery for a modal epistemic system we can rely on a set-theoretical structure that includes, at the very least, a set *W* of possible worlds and an "accessibility relation" *R* among worlds. Intuitively, if w_1 and w_2 are two worlds linked by *R*, then the agent cannot discriminate among them (i.e., for all he knows, if he is w_1 he might well be in w_2). The validity of certain axioms rather than others depends on the structure of the accessibility relation between worlds. Thus, for example, if *R* is reflexive (i.e., if each world can be related to itself) we guarantee that **T** holds, whereas the transitivity of *R* is necessary and sufficient for ***KK***.

In "Improbable Knowing",[11] Williamson considers a frame $<W, R, P_{prior}>$ for a single agent, where *W* is a set of worlds, *R* is an accessibility relation between worlds, and P_{prior} is a prior probability distribution defined over subsets of *W*. It is assumed that *W* is finite and that P_{prior} is uniform, in order not to add useless complications.[12]

As usual, propositions are subsets of *W*, and, if φ is a proposition, then *K*φ is the set of all worlds connected with φ-worlds through *R*:

[11] Williamson (mn).

[12] Hence P_{prior} is *regular*, in the sense that $P_{prior}(\varphi) = 0$ iff $\varphi = \varnothing$.Williamson takes priors in his system to refer to the intrinsic plausibility of worlds prior to our gathering any evidence (cf. also his (2000), chapters 9 and 10). If we feel uncomfortable with this extremely objectivist picture, we can always take priors to embody the personal measures of the theoretician –who can in turn be conceptualized as the subject who seeks to make knowledge attributions to third party agents.

$$K\varphi = \{w \in W: \forall x \in W\,(wRx \rightarrow x \in \varphi)\}$$

Define also $R(w)$ as the strongest proposition known in w:

$$R(w) = \{x \in W: wRx\}$$

It is easy to see that, by definition of $K\varphi$, $R(w)$ is included in every proposition known by the agent. On the other hand, given that, by hypothesis, P_{prior} is a uniform probability measure and W is finite, $P_{prior}(\varphi)$ will just amount to $\#[\varphi] / \#W$. Williamson also defines φ's probability in a given world w, or φ's *evidential probability* in w, which shall be written as $P_w(\varphi)$. $P_w(\varphi)$ is obtained by conditionalizing on what the agent knows in w, i.e., on $R(w)$. Hence,

$$P_w(\varphi) = P_{prior}(\varphi \mid R(w)) = P_{prior}\,(\varphi \cap R(w)) / P_{prior}\,(R(w))$$

This definition is in agreement with the much discussed E=K thesis, according to which the agent's evidence at a particular moment is no less than the totality of his or her knowledge. As we have a uniform prior distribution, in order to calculate φ's evidential probability in w we consider how many of the $R(w)$-worlds are also φ-worlds. A natural consequence of this idea is that, for all w, $P_w(R(w)) = 1$.

We can also wonder about the extension of a proposition stating that φ's probability is r; Williamson defines it as the set of worlds in which φ has evidential probability r:

$$[P(\varphi) = r] =_{\text{def.}} \{w \in W: P_w(\varphi) = r\}$$

Within this setting, Williamson suggests that the ***KK*** principle can be formulated in probabilistic terms: "The ***KK*** principle is equivalent to the principle that if the evidential probability of p is 1, then the evidential probability that the evidential probability of p is 1 is itself 1" (Williamson, ms, p.8). It is easy to show that this claim is false if R is not transitive; in fact, when R is not transitive we can propose examples in which $P_w([P(R(w))=1])$ is as low as we want.

4. Second Thoughts about Second-Order Probabilities

According to the intended interpretation of Williamson's proposal, whenever we assess $[P(\varphi) = r]$'s probability, for any φ, we are actually assessing a *higher-order* probability. However, I believe this claim is problematic. The expression between square brackets is

just a label to refer to a certain set of worlds; thus, at the time of assessing the probability of "[$P(\varphi) = r$]", it would not make any difference if we write it as, say, "ψ", without any reference to probabilities whatsoever – as long as the set of worlds remains the same. So there is a sense in which in such a case we might just as well be calculating a *first*-order probability. The root of the problem, I take it, is that propositions understood as sets of worlds are too coarse-grained to allow for what we want: second-order probabilities call for a more fine-grained representation device. (As we shall see, the reason is ultimately that second-order probabilities demand that we conditionalize over second-order evidence.) Thus, even if "φ" and "ψ" are logically equivalent (in the sense that they refer to the same set of worlds in W), their probability might differ. In light of this, in what follows I will suggest a representation strategy that enable us to take into account not only propositions understood as sets of worlds, but also their *mode of presentation*, so to speak.[13]

Let me then suggest a model in which genuine second-order probabilities apply to well-regimented probabilistic statements, rather than to sets of worlds. Thus the probability of a set of worlds (or proposition) will depend crucially on the way we refer to it – in particular, on whether we refer to it through a probabilistic discourse or not.[14] We will then take the arguments of probability functions of our system to be sentences of a sequence of duly regimented languages. As usual, if $\underline{\varphi^i}$ is a formula of L^i, $[\varphi^i]$ will be the proposition, or set of worlds, in which $\underline{\varphi^i}$ is true; I will eliminate the square brackets when no risk of confusion arises. In what follows, sentences and other linguistic items will always appear as underlined expressions.

To carry out this project we need to enrich the original frame with a function v that helps us assess the truth-values of sentences of a sequence of languages L^0, $L^1 \ldots L^n \ldots$, with probability operators $\underline{\mathbf{P}^0}$, $\underline{\mathbf{P}^1} \ldots \underline{\mathbf{P}^n} \ldots$ of increasingly higher levels. We will consider also a sequence of functions $P^1_w \ldots P^n_w \ldots$ (for each $w \in W$), which will be

[13] We might rather choose to refine W and allow for metaphysically impossible worlds (say, worlds in which "φ" and "ψ" are neither both true nor both false (thanks to Timothy Williamson for this point). But the approach adopted in this paper seems more natural, and respects the intuition that purely linguistic differences sometimes matter, even to fully rational agents.

[14] For other well-known frameworks that deal with higher-order probability, cf. Skyrms (1980), Gaifman (1986), Samet (1997), or van Fraassen (1995).

applied to increasingly complex arguments. To put it somewhat sloppily, in each case P^i_w will take as arguments sentences of language *i*-1:

$P^i_w: L^{i-1} \rightarrow \mathbb{R}$

(A more careful presentation, as well as further details on language formations rules, will be given in section 5). Expressions of the form $P_{prior}(\underline{\varphi})$ or $P^i_w(\underline{\varphi})$ will not be part of any language of the sequence $L^0, L^1 \ldots L^n \ldots$, but they will belong to the metatheory. In this way we make sure we are not mixing up truths of the system with truths *in* the system.

Following Williamson's proposal, prior probability will amount to the cardinality of the set of worlds in which a sentence of some language is true, given function *v*, divided by the cardinality of *W*. On the other hand, it is natural to demand that, for all *w*, $\underline{\mathbf{P}^i(\varphi)=r}$ be true in *w* iff $P^i_w(\underline{\varphi}) = r$, where "$\underline{\mathbf{P}^i(\varphi)=r}$" is a sentence of L^i, and "$\underline{\varphi}$" is a sentence of L^{i-1}. The central problem now is how to define evidential probability *i* in a world – in other words, how to conditionalize.

Suppose we have information about the state of the weather tomorrow. We have read the forecast in the newspaper, watched the weather channel, etc. On the basis of all this, we conclude that the probability of rain tomorrow is *r*. Now suppose a friend asks us how probable it is that our rational degree of belief that there's rain tomorrow is in fact *r*. As I see it, in this case our friend is no longer interested in the probability of a proposition about meteorology, but in the probability of a proposition *about the degree of confirmation* possessed by our original meteorological statement. Which is the relevant evidence to answer this question, then? Intuitively, what we have to assess is how good we are at the time of engaging in confirmation theory. Thus the relevant total evidence is no longer *R*(*w*), but a *second-order* evidence: the evidence for our second-order probability should consist in what we know about our capabilities to adequately confirm propositions. And the strongest proposition that expresses this idea is indeed *KR*(*w*). Notice, incidentally, that the idea of conditionalizing on higher-order evidence corpora when dealing with second-order knowledge can be taken to be in perfect agreement with the K=E thesis, well understood.

How should we conditionalize, then? A first suggestion could be to apply the following rule:

For $i \geq 1$: $P^i_w(\underline{\mathbf{P}^{i-1}(\ldots\mathbf{P}(\varphi)=r\ldots)}) =$

$P_{\text{prior}}(\underline{\mathbf{P}^{i-1}(\ldots\mathbf{P}(\varphi)=r\ldots)} \,\&\, \underline{K^{i-1}\ldots KR(w)}) \,/\, P_{\text{prior}}(\underline{K^{i-1}\ldots KR(w)})$

Here the sequence of $\underline{K}$s is simply the result of iterating the *same* $\underline{K}$ operator as many times as $\underline{\mathbf{P}}$-operators are in the nominator's argument. The language level to which the argument belongs (as indicated by $\mathbf{P}$'s super-index), determines the order of the probability function whose value we are seeking to calculate, and fixes the number of $\underline{K}$s we'll have to iterate to make the calculation.

This first proposal is not completely satisfactory, though, because it is easy to show that it leads us to divorcing probability 1 from knowledge: there will be models in which $P^2_w(\underline{\mathbf{P}^1(R(w))=1}) = 1$ and yet $\underline{KKR(w)}$ is not true in w.[15]

To overcome this difficulty, we can enrich languages L^0, $L^1 \ldots L^n \ldots$with a sequence of knowledge operators $\underline{K^0}$, $\underline{K^1} \ldots \underline{K^n} \ldots$ that runs parallel to our sequence of *probability* operators. We will need, then, a family of relations $R^1 \ldots R^n \ldots$ for $\underline{K^1} \ldots \underline{K^n} \ldots$ Thus, the desire that probability and knowledge claims cohere with each other motivate us to propose a model with multiple *K*-operators. I will discuss the legitimacy of this motivation with some detail in section 6. But before that, let me describe the formal proposal more carefully.

5. A Model for Temperate Transparency

Consider then the model $\mathcal{M} = \langle W, R^1, \ldots R^n \ldots, P_{prior}, v \rangle$, where:

(1) W is a (finite) set of worlds.[16]

(2) R^i is a reflexive relation over W, for all i, and transitive for $i > 1$.

(3) $R^i \subseteq R^{i-1} \ldots \subseteq R^1$

[In the next section we will discuss a rationale for demanding (2) and (3), as well as further possible requirements.]

(4) There is a sequence of languages $L^0, L^1 \ldots L^n \ldots$ such that:

[15] Suppose $R = \{(w, w), (w, x), (x, x), (y, y), (y, y)\}$. Then $[R(w)] = \{w, x\}$, $[KR(w)] = \{w\}$, and $[KKR(w)] = \varnothing$. However, $P^2_w(\underline{P(R(w))=1}) = 1$.

[16] I follow Williamson in demanding that W be finite, but this restriction can of course be abandoned – in which case some of the clauses that follow would need to be appropriately amended.

a) $\underline{p_1}...\underline{p_n}$ are atomic formulas of L^0. Atomic formulas are well formed formulas (wff).

b) If $\underline{\varphi}$ is a wff of L^i, $\underline{\varphi}$ is a wff of L^{i+n}, for any n.

c) If $\underline{\varphi}$, $\underline{\psi}$ are wff of L^i, so are $\underline{\sim\varphi}$, $\underline{\varphi\vee\psi}$, for any $i \geq 0$. (And, as usual, we have $\underline{\varphi\rightarrow\psi} =_{df} \underline{\sim\varphi\vee\psi}$, and $\underline{\varphi\&\psi} =_{df} \underline{\sim(\sim\varphi\vee\sim\psi)}$.

d) If $\underline{\varphi}$ is a wff of L^0, $\underline{K^1\varphi}$ belongs to the K-fragment of L^1. Formulas in the K-fragment of L^i are wff of L^i, for any $i \geq 0$.

e) If $\underline{\varphi}$ belongs to the K-fragment of L^i, so does $\underline{\sim\varphi}$. Nothing else belongs to the K-fragment of L^i.

f) If $\underline{\varphi}$ belongs to the K-fragment of L^i, $\underline{K^{i+1}\varphi}$ belongs to the K-fragment of L^{i+1} (for $i \geq 1$).

g) If $\underline{\varphi}$, $\underline{\psi}$ are wff of L^0: $\underline{\mathbf{P}^1(\varphi)=r}$, $\underline{\mathbf{P}^1(\varphi|\psi)=s}$ belong to the P-fragment of L^1 (for any r, s in [0,1]). Formulas in the P-fragment of L^i are wff of L^i, for any $i \geq 1$.

h) If $\underline{\varphi}$, $\underline{\psi}$, belong to the P-fragment of L^i, so do $\underline{\sim\varphi}$, $\underline{\varphi\vee\psi}$. Nothing else belongs to the P-fragment of L^i.

i) If $\underline{\varphi}$ is in the P-fragment of L^i, then $\underline{\mathbf{P}^{i+1}(\varphi)=r}$ belongs to the P-fragment of L^{i+1}.

j) If $\underline{\varphi}$, $\underline{\psi}$ are wff of L^i, and either $\underline{\varphi}$ or $\underline{\psi}$ belongs to the P-fragment of L^i, then $\underline{\mathbf{P}^{i+1}(\varphi \mid \psi)=r}$ belongs to the P-fragment of L^{i+1}.

k) Nothing else is a wff of L^0, $L^1 \ldots L^n \ldots$

To keep with the spirit of our prior terminology, at times it will be convenient to use "$\underline{R(w)}$" (for $w \in W$) as a shortcut for the relevant wff of L^0, such that for any $w \in W$, "$\underline{R(w)}$" is true in all worlds x such that wR^1x. Also, if "$\underline{\varphi}$" is a sentence of L^0, $L^1 \ldots L^n \ldots$, then $[\varphi]$ is the set of worlds in which $\underline{\varphi}$ is true.

(5) v is a function that maps atomic formulas of L^0 into sets of worlds.

Then the assessment of sentences in the model follows the usual pattern:

$\vDash_w \underline{p_i}$ iff $w \in v(p_i)$

$\vDash_w \underline{\sim\varphi}$ iff not-$\vDash_w \underline{\varphi}$

$\vDash_w \underline{\varphi\vee\psi}$ iff either $\vDash_w \underline{\varphi}$ or $\vDash_w \underline{\psi}$

$\vDash_w \underline{K^i\varphi}$ iff $\forall x \in W$: if wR^ix, then $\vDash_x \underline{\varphi}$[17]

$\vDash_w \underline{\mathbf{P}^i(\varphi)=r}$ iff $P^i_w(\underline{\varphi}) = r$ [18]

$\vDash_w \underline{\mathbf{P}^i(\varphi \mid \psi)=r}$ iff $P^i_w(\underline{\varphi} \mid \underline{\psi}) = r$ [19]

(6) $P_{prior}(\text{-})$ is a probability function on sentences of $L^0, L^1 \ldots L^n \ldots$, and $P_{prior}(\text{-}|\text{-})$ is a conditional probability function on pairs of sentences of $L^0, L^1 \ldots L^n \ldots$, such that

1. $P_{prior}(\underline{\varphi}) = \#\{w: \vDash_w \underline{\varphi}\} / \#W$; and
2. $P_{prior}(\underline{\varphi} \mid \underline{\psi}) = \#\{w: \vDash_w \underline{\varphi} \ \& \vDash_w \underline{\psi}\} / \#\{w: \vDash_w \underline{\psi}\}$

(7) $P^i_w(\underline{\varphi})$ is an unconditional probability function on the *P*-fragment of $L^{i-1,}$ such that $P^i_w(\underline{\varphi}) = P_{\text{prior}}(\underline{\varphi} \mid \underline{K^{i-1}...KR(w)})$.

(8) $P^i_w(\underline{\varphi} \mid \underline{\psi})$ is a conditional probability function, where both $\underline{\varphi}$ and $\underline{\psi}$ belong to L^{i-1}, and at least one of them belongs to the *P*-fragment of L^{i-1}, such that $P^i_w(\underline{\varphi} \mid \underline{\psi}) = P_{\text{prior}}(\underline{\varphi} \mid \underline{\psi} \ \& \ \underline{K^{i-1}...KR(w)})$.[20]

A few comments are in order. To simplify, at times we will simply rely on R^+ to refer to higher-order R^is (for $i > 1$). In addition, let $\mathscr{P}^i$ be the set of wff of L^i; let $\mathscr{P}_K{}^i$ be the set of wff of the *K*-fragment of L^i, and let $\mathscr{P}_P{}^i$ the set of wff of the *P*-fragment of L^i. We then have $\mathscr{P}^i = Cn(\mathscr{P}^{i-1} \cup \mathscr{P}_K{}^i \cup \mathscr{P}_P{}^i)$, as well as $\mathscr{P}^0 \subset \mathscr{P}^1 \subset \ldots$ Notice that $\underline{K^i}$: $\mathscr{P}_K{}^{i-1} \rightarrow \mathscr{P}_K{}^i$ (for $i \geq 1$), so $\underline{K^i}$ is not strictly speaking an "operator" and $\mathscr{P}_K{}^i$ is not closed under Boolean connectives. This is, I think, as it should be, considering the intended meaning of the formalism (see below). In any case, to keep the terminology simple, I will

[17] In other words, for all *i*, $[K^i\varphi] = \{y \in W: \forall x \in W\ (yR^ix \rightarrow x \in \varphi)\}$. Notice that, for $i > 1$, "$\underline{K^i\varphi}$" is a wff only if "$\underline{\varphi}$" is of the form "$\underline{K^{i-1}\psi}$" or "$\underline{\sim K^{i-1}\psi}$", in agreement with 4.d, 4.e and 4.f.

[18] Notice that "$\underline{\mathbf{P}^i(\varphi) = r}$" is well formed only if "$\underline{\varphi}$" belongs to the *P*-fragment of L^{i-1}.

[19] Notice that "$\underline{\mathbf{P}^i(\varphi \mid \psi) = r}$" is well formed only if both "$\underline{\psi}$" and "$\underline{\varphi}$" belongs to L^{i-1}, and either "$\underline{\psi}$" or "$\underline{\varphi}$" belongs to the *P*-fragment of L^{i-1}.

[20] This flexibility regarding the nature of the arguments of conditional evidential functions will enable us to establish some important links between lower- and higher-order probabilities, as we shall see in section 8.

continue to refer to the sequence of K^is as a sequence of knowledge operators, in a loose sense. A similar point applies to probabilistic sentences within the sequence of languages in the model. We actually have $\underline{\mathbf{P}^i}(\text{-})$: $\mathscr{P}_P^{i-1} \rightarrow \mathscr{P}_P^{i}$, as well as $\underline{\mathbf{P}^i}(\text{-}|\text{-})$: $(\mathscr{P}_P^{i-1} \times \mathscr{P}^{i-1}) \cup (\mathscr{P}^{i-1} \times \mathscr{P}_P^{i-1}) \rightarrow \mathscr{P}_P^{i}$. As with their knowledge counterparts, I will speak loosely of "probability operators" to refer to the $\underline{\mathbf{P}^i}$s.

Let me stress that, according to the intended interpretation, an expression such as "*S* has second-order knowledge that *p*" (i.e., (*)"$\underline{K^2 p}$") does not make sense. To have second-order knowledge means that, *on reflecting on our beliefs*, we find to know or not to know that something is the case. Hence it is just appropriate to require that "$\underline{K^2}$" always be followed by a first-order operator or its negation.[21] The interpretation of higher-order levels of knowledge follows the same spirit (thus, for example, the intended interpretation of "$\underline{K^3}$" has it that, on reflecting on our reflection about our beliefs, we find to know or not to know that we know or do not know that something is the case, and so on). This is in strike contrast with the intended meaning of first-order knowledge. To have first-order knowledge that *p* (i.e., "$\underline{K^1 p}$") means that, on reflecting *on the world*, we find it to be the case that *p*. Incidentally, note that the attitude we take towards ignorance in the first-order case is typically very different from the attitude we take towards ignorance in higher-order levels. In general, if the considerations of section 2 were on the right track, ideal agents can be assumed to be aware of their own epistemic states, whereas they are not assumed to be empirically omniscient. This reinforces the motivation for having different knowledge operators.

We can of course discuss how much higher up in the hierarchy actual agents are able to grasp well-formed sentences. This is an empirical question, and one we should not worry about in this context. Clearly, we should not put *a priori* limitations to the levels agents could reach on careful reflection.

The present framework validates *Modus Ponens* and *Generalized Necessitation*: if $\vdash \underline{K^i \varphi}$, then $\vdash \underline{K^{i+1} K^i \varphi}$ (for any $i \geq 0$), as well as the following axioms:

[21] Related to this, notice that, even though "$\underline{q \& Kq}$" is expressible in L^2 (by 4.b and 4.c), "$\underline{K^2(q \& Kq)}$" is not. This should not be counterintuitive, once we consider what a second-order knowledge operator means in the model. It could be objected that agents sometimes do express sentences such as "I know *both* that the train has just arrived and that I know it has come from abroad". But it can well be argued that in such cases what the agent actually wants to convey is *both* that she knows that the train has just arrived, *and* that she has second-order knowledge that she knows that the train has come from abroad. The correct way of rendering this expression would then be "$\underline{Kq \ \& \ K^2 Kq}$", which is a perfectly well-formed formula in the model. Similar remarks hold for cases in which the agent considers the knowledge possessed by third party agents (which is actually akin to having knowledge *of the world*), and compares it with her own knowledge.

- **K**: $\underline{K^i(\varphi \rightarrow \psi) \rightarrow (K^i\varphi \rightarrow K^i\psi)}$ [Notice that $\underline{K^i(\varphi \rightarrow \psi)}$, $\underline{K^i\varphi}$ and $\underline{K^i\psi}$ are wff only if i=1]
- **Generalized T**: $\underline{K^i\varphi \rightarrow \varphi}$

  [for any $i \geq 1$, and $\varphi \in$ the K-fragment of L^{i-1}.]
- **KK+**: $\underline{K^i\varphi \rightarrow K^{i+1}K^i\varphi}$ [for $i > 1$]

(See the Appendix)

As for the intended meaning of higher-order probabilities, recall that, as opposed to *first*-order probabilities, an evidential probability claim of second-order degree is the evidential probability *of a probability statement* (a wff of a *P*-fragment of a language in the model). A conditional evidential probability of second-order degree, on the other hand, is either (i) the evidential probability of a well formed (first-order) probability statement conditional on another well formed probability statement, or (ii) the evidential probability of any well formed statement conditional on a probability statement, or perhaps (iii) the evidential probability of a probability statement conditional on another well formed statement that may or may not be itself probabilistic. Hence the restrictions on the arguments, as found in requirements (7) and (8). Recall that, in the present proposal, the (meta-theoretic) statements we can prove *about* the model are not among the statements expressible by languages *in* the model. Thus probability functions that help us express truths about the model (say, from the theoretician's perspective) do not conflate with probability operators of $L^1 \ldots L^n \ldots$

As it should be clear from (7) and (8), our conditionalization rule will now incorporate sentences with operators $\underline{K^1}, \ldots \underline{K^n} \ldots$ whose behavior is regulated by $R^1, \ldots R^n \ldots$ To put it more explicitly, we will have

For $i \geq 1$: $P^i_w(\underline{\mathbf{P}^{i-1}(\ldots \mathbf{P}(\varphi)=r\ldots)}) =$

$P_{\text{prior}}(\underline{\mathbf{P}^{i-1}(\ldots \mathbf{P}(\varphi)=r\ldots)} \,\&\, \underline{K^{i-1}\ldots KR(w)}) \,/\, P_{\text{prior}}(\underline{K^{i-1}\ldots KR(w)})$

As opposed to the first proposal (in section 4), the relevant sentences can no longer contain iterations of the same $\underline{K}$ operator, but they will include i-1 higher-order $\underline{K}$-operators. The language level to which the argument belongs (as indicated by $\underline{\mathbf{P}}$'s super-index), determines the order of the evidential probability function whose value we are seeking to calculate, and fixes $\underline{K^{i-1}}$'s degree. *Mutatis mutandis* for conditional evidential probability.

Depending on how R^1 is, it may be important to incorporate additional requirements for R^+, in order to satisfy the self-imposed demand that our attributions of probability and knowledge coexist in a coherent way; intuitively, different restrictions will correspond to different degrees of idealization of the epistemic agent involved. Thus, for instance, we could demand that R^2 be such that:

(a) $\forall w \forall x \in W\,(wR^2x \rightarrow x \in [KR(w)])$

(i.e., if wR^2x, then for any y such that xR^1y, we also have wR^1y). Notice that there might be more than one way to satisfy this requirement. The most conservative way to comply with it would be to strengthen (a) to a biconditional, in which case $[K^2KR(w)] = [KR(w)]$, for any w. Then R^2 differs from R^1 as little as possible without violating (a). For the least conservative way to satisfy (a), just let R^2 be the identity relation (*Id*):

(b) $\forall w \forall x \in W\,(wR^2x \rightarrow w{=}x)$,

in which case $[K^2KR(w)]$ might be a proper subset of $[KR(w)]$, for some w.

More modestly, we could rather demand that R^2 satisfy:

(c) $\exists w \forall x \in W\,(wR^2x \rightarrow x \in [KR(w)])$

(i.e., we could demand that $[K^2KR(w)] \subseteq [KR(w)]$, for some w). I will discuss a rationale for these requirements in the next section.

6. Discussion

Let me discuss some of the consequences that follow from demanding more or less strict requirements for our sequence of *R*s. First of all, I would like to address a prior worry on the structure of R^1. Someone might wonder at this point why not demand that R^1 be an equivalence relation, and avoid any further complication. But we do not want to impose such a restrictive structure on R^1. Indeed, the failure of transitivity for R^1 seems just as appropriate, given, among other things, Williamson's convincing considerations on margin of error principles for knowledge (cf. section 1). The fact that I know that p in a close world w_1 need not mean that I still know that p in a world w_n that is utterly different from the actual one, only by virtue of there being a chain of words between w_1 and w_n, any of which differs from its neighbors only slightly. In other words, "old fashioned" violations of the (unqualified) ***KK*** Principle seem very well-motivated to me.

Now, by demanding nested *R*s in higher levels we may lose ordered pairs while we go up (we let worlds to be progressively more "isolated", so to speak). By demanding transitivity, moreover, we ensure that for higher-order levels we will have an analogous to ***KK*** for successive operators, as we shall see. I would like to stress here that these demands are not *ad hoc*:

I have argued that, in order to calculate a second-order evidential probability at world *w*, the agent should conditionalize over $\underline{KR(w)}$, rather than over $\underline{R(w)}$. To make this move possible, I have suggested that genuine higher-order probabilities apply to probabilistic statements rather than to sets of worlds. This idea led us to define a sequence of languages with increasingly complex probabilistic claims. As a result of this strategy we obtain that the second-order probability claim stating that the probability of $\underline{R(w)}$ is 1 is also 1. Thus, in order to have knowledge and probability concepts that fit with each other we should also say that the agent *knows* that $\underline{KR(w)}$. This, in turn, forces us to define a sequence of knowledge operators.[22] To put it briefly, we want to have a sequence of *K*s that reflects, with a non-probabilistic vocabulary, the idea that we have probability 1 at higher-order levels (when we do have it). The choice of the structure of the *R*s is then the result of seeking that probability and knowledge claims complement each other in a coherent way.

So far I have been assuming that, if *S*'s total knowledge allows *S* to give probability 1 to a particular statement (perhaps even a probabilistic statement), then we should be entitled to say that *S knows* the truth of that particular statement. This is indeed the crucial assumption that motivate us to define a particular system of knowledge operators, and which will eventually lead to the vindication of transparency principles of some sort. Many philosophers, however, have been reluctant to accept this assumption, mostly for reasons related to the nature of infinite domains, where probability 1 is not certainty. If *p*'s probability can be 1 and still *p* be false, then the inference from probability 1 to knowledge should surely fail. Or so the objection goes.

The objection, however, needs to be seriously qualified. What the objection actually does is provide us with a strong reason to distinguish between different types of probability 1 in infinite models. For an infinite *W*, $P^1_w(\underline{\varphi}) = 1$ should not always force

[22] Philosophically speaking, the fact that we end up with different *K*-operators can be seen to reinforce the intuition that, when an agent reflects on her mental states, she is not dealing with the same type of phenomenon as when, say, she sees a tree in front of her. Thus, it is just appropriate to have "$\underline{K^2KR(w)}$", rather than (*)"$\underline{KKR(w)}$" (the latter is no longer a legitimate formula of the system).

the agent to have $\models_w \underline{K\varphi}$, *but sometimes this is indeed required*, namely, when $[R(w)] \subseteq [\varphi]$. Hence, were we working with an infinite W, it would be advisable to make a distinction between cases of $P^1{}_w(\varphi) = 1$ in which $[R(w)] \subseteq [\varphi]$, and those in which $[R(w)] \not\subset [\varphi]$; *mutatis mutandis*, this observation applies to higher-order levels as well.

Therefore, in an infinite frame the motivation for allowing for multiple K-operators still holds. Only, when W is infinite we should no longer argue from $\models_w \underline{\mathbf{P}^{\mathbf{i+1}}(\ldots(\mathbf{P}^{\mathbf{1}}(\varphi)=1)\ldots)=1}$ to $\models_w \underline{K^{i+1}\ldots K\varphi}$, but from $[K^i\ldots KR(w)] \subseteq [K^i\ldots K\varphi]$ (and hence $\models_w \underline{\mathbf{P}^{\mathbf{i+1}}(\ldots(\mathbf{P}^{\mathbf{1}}(\varphi)=1)\ldots)=1)}$ to $\models_w \underline{K^{i+1}\ldots K\varphi}$. Here I will not make further comments on how the structure of such an infinite frame might go; a more detailed account will be left for further work. But, for the meantime, these remarks should suffice to appease some worries.

Are requirements (1) to (8) from section 6 enough to secure that the model has all the consequences we would like it to have? Not quite; properties (a), (b) or (c) may be important as well.

If R^2 satisfies property (a) from section 6, we guarantee that $[K^2KR(w)]$ will not be empty, for any w, regardless of the structure of the original R. This, in turn, guarantees the existence of even higher order evidential probabilities for world w, which will be obtained by conditionalization on $\underline{K^2KR(w)}$. From a philosophical point of view, the resulting system can be said to describe *an ideally responsible agent*, insofar as ideally responsible agents are assumed to know all they know.[23] It is easy to see that in such a system we obtain the validity of principle $\underline{K\varphi \rightarrow K^2K\varphi}$ for all $\underline{K\varphi}$ in L^1, which we might dub ***the KK^2 Principle***. (See the Appendix; indeed, we also obtain its generalization for higher levels). On the other hand, $\underline{K\varphi \rightarrow K^2K\varphi}$ is not valid in $\mathcal{M}$ if property (a) is violated.

Moreover, when R^2 is the identity relation (i.e., if it fulfills property (b)), there is an even stricter demand on the agent's introspective capabilities; it can be seen as the analogous of demanding full logical omniscience for the representation of ideally *rational* agents. $R^2 = Id$ will become important at the time of considering conditional

[23] Here we have a further reason why "$\underline{K^2p}$" cannot be a wff of a language in the model. Assuming S is ideally responsible of her own intentional states, she can be asked to know all she knows, but she cannot be sensibly asked to know all that is true.

evidential probabilities; thus, by fulfilling property (b) we have an ideal fit between knowledge and probability claims, in a sense to be discussed in the next section.

When R^2 does not satisfy (a) (hence it does not satisfy (b) either), we relax the coherence demand imposed on our system, because there could be a world w such that $P^2_w(\underline{\mathbf{P}^1(\varphi)=r}) = 1$ and yet not-$\vDash_w \underline{K^2KR(w)}$. As is obvious, if $\underline{K^2KR(w)}$ is not true in w, P^3_w will not be defined. We can interpret this result as evidence that it will not always be meaningful to keep on going "higher" in our probability assignments (say, because our epistemic capabilities are limited). When higher-order probabilities are undefined in w, we will say that *the agent does not have responsible knowledge of* $\underline{R(w)}$.

A system in which the agent does not have responsible knowledge in *any* world is a system in which the coherence between probabilities and knowledge is extremely poor. Requirement (c) from section 6 is enough to guarantee that there is at least one world in which probability and knowledge claims do not part ways. In such a world the agent's beliefs are "in the region of responsibility". Were we to enrich a system of this type with **S5** alethic operators, we would obtain the validity of $\underline{\Diamond (K\varphi \rightarrow K^2K\varphi)}$. Let us call it ***the KK$^{\Diamond}$ Principle.***

Even if we do not add further requirements beyond (1) to (8), the transitivity of R^+ ensures the validity of $\underline{K^2K\varphi \rightarrow K^3K^2K\varphi}$, for all $\underline{K\varphi}$, as well as the more general $\underline{K^i\ldots K\varphi \rightarrow K^{i+1}K^i\ldots K\varphi}$, for $i > 1$. We will call it ***the KK$^+$ Principle***. We can also prove that $P^2_w(\underline{\mathbf{P}^1(\varphi)=r})$ will be always 1 when r is either 0 or 1, whereas (unconditional) evidential probability at level 2 will be guaranteed to be 1 or 0 if R^1 is an equivalence relation. It is interesting to notice that if the evidential probability at level 1 for an arbitrary proposition [φ] is r, for $0 \neq r \neq 1$ and R^1 not transitive, then the corresponding evidential probability at level 2 (*i.e.*, the evidential probability that the first level probability is r) need not be 1, which means that going up in the hierarchy will not always be trivial (see the Appendix).

We shall say that ***KK$^+$, KK$^{\Diamond}$*** , and ***KK^2*** **are quasi-transparency principles,** which describe different degrees of idealization we can demand from agents. Notice that the validity of ***KK$^+$, KK$^{\Diamond}$*** and ***KK^2***, in each case, was not imposed from the outside, as it were, but was obtained as a consequence of the natural injunction to conditionalize over increasingly higher orders of evidence, while at the same time attempting to adjust probability-language and knowledge-language in a progressively coherent way.

7. Further Consequences. A Note on Probabilistic Reflection

Some of the results highlighted in section 6 referred to the way lower- and higher-level unconditional probabilities relate to each other in the model. In this section I will discuss briefly some ways to link lower and higher-level *conditional* probabilities. In particular, we may wonder about the correction of the so-called *Reflection Principle* in probability. To distinguish it from ***KK*** let me refer to it as *the Probabilistic Reflection Principle*, or PRP. A standard formulation of PRP goes as follows:

$$P(\alpha \mid P(\alpha) = r) = r$$

(where "α" is a proposition and $r \in [0, 1]$). There are many possible interpretations of the principle in the literature, which shall not be discussed here.[24]

Consider, first, how to translate it to our present notation. Clearly, the relevant probability function should be (at least) a second-level function. Thus, PRP has it that the (second-order) evidential conditional probability of a sentence $\underline{\varphi}$ of L^0 in a given world *w*, given the truth of the sentence stating that the probability of $\underline{\varphi}$ is *r*, is itself *r*:

$$P^2_w(\underline{\varphi} \mid \underline{\mathbf{P}^1(\varphi)=r}) = r \text{ (for } w \in W),$$

which, in turn, should be rendered as:

$$P_{prior}(\underline{\varphi} \mid \underline{\mathbf{P}^1(\varphi)=r \;\&\; KR(w)}) = r$$

Moreover, our present framework enables us to formulate PRP for higher levels, such as:

$$P^3_w(\underline{\varphi} \mid \underline{\mathbf{P}^2(\varphi \mid \mathbf{P}^1(\varphi)=r)=r}) = r,$$

as well as the more general *Iterated PRP*:

$$P^i_w(\underline{\varphi} \mid \underline{\mathbf{P}^{i-1}(\varphi \mid \mathbf{P}^{i-2}(\varphi \mid \ldots)\ldots.)=r}) = r,$$

or, equivalently:

$$P_{prior}(\underline{\varphi} \mid \underline{\mathbf{P}^{i-1}(\varphi \mid \mathbf{P}^{i-2}(\varphi \mid \ldots)\ldots. = r \;\&\; K^{i-1}\ldots KR(w)}) = r$$

We will say that Iterated PRP is a theoretical truth of a model $\mathcal{M}$ that satisfies requirements (1) to (8) from section 6 iff for every *w* in *W* in which P^i_w exists, $P^i_w(\underline{\varphi} \mid \underline{\mathbf{P}^{i-1}(\varphi \mid \mathbf{P}^{i-2}(\varphi \mid \ldots)\ldots.) = r}) = r$, for any $i \geq 2$. More flexible combinations can be discussed as well. For example, Iterated PRP could hold for upper levels only (say, for *i*

[24] The interested reader is referred to Skyrms (1980).

$\geq n \geq 3$), or just for some worlds of *W*. In case it is true for some worlds and not for others, we may label such worlds as particularly desirable, from an epistemic point of view – as we did in the previous section with epistemically optimal worlds in which agents are ideally responsible.

Is *Iterated PRP* a theoretical truth of $\circlearrowright\mathcal{M}$? Or, at the very least, does it hold for some $i \geq 3$? And, in case the answer is negative, how bad is this result? Clearly, the truth of PRP depends on the structure of the *R*s. Without additional requirements, a model fulfilling claims (1) to (8) does not make PRP true, and cannot guarantee the truth of Iterated PRP, for any finite *i*. However, with a few additional demands – some of them already discussed in the previous section – some version of the principle can be secured, as we shall see.

In any case, let me point out first that it is not clear to me whether we are entitled to ask that PRP be satisfied *when working with evidential probabilities*. Recall that "the evidential probability that *p*", for an agent *S*, is actually rendered as: "the probability that *p*, given all *S* knows". Then, "the evidential probability that *p*, given the truth of the sentence stating that the probability that *p* is *r*" is equally rendered as "the probability that *p*, given that $\underline{\mathbf{P}^1(p)=r}$ *and* given that, on reflecting on her beliefs, *S* finds it to be the case that she knows that….". But the second conjunct may well affect how confident *S* is in "$\underline{\mathbf{P}^1(p)=r}$". Thus, PRP should only hold when *S*'s knowledge does not have a chance to affect *S*'s confidence in the truth of "$\underline{\mathbf{P}^1(p)=r}$". But, assuming all *S* knows is $R(w)$, this will be the case only when $P^2_w(\underline{\mathbf{P}^1(p)=r}) = 1$; as we have seen in the previous section, this means, in turn, either that $r = 1$ or that R^1 is an equivalence relation.

As it happens, Williamson has proved that "for a regular probability distribution over a finite serial frame, the reflection principle holds iff the frame is quasi-reflexive, quasi-symmetric and transitive."[25] As R^1 is assumed to be reflexive (due to the factivity of knowledge), this amounts to saying that R^1 should be an equivalence relation (seriality, in turn, is meant to ensure that no evidential probabilities at the lower level go undefined). In light of our discussion from the previous paragraph, this comes as no surprise: it is exactly as it should be. I do not think, however, that we should demand symmetry and transitivity at the lower level *just in order to guarantee that PRP holds* – it should be clear that violations of the principle motivated by the fact that *S*'s

[25] Corollary 5, informal communication. Similar results hold for a regular countable additive probability distribution over a serial frame (Corollary 7). Thanks to Horacio Arló-Costa for pointing out these results to me.

knowledge affects *S*'s confidence in $\underline{\mathbf{P}^{\mathbf{i}}(p)=r}$ need not be a symptom of *S*'s irrationality. Moreover, I have already argued why it is not sensible to demand that, for every well-behaved epistemic model, R^1 be an equivalence relation.

At higher levels, however, other considerations become important, as we have seen. So perhaps violations of Iterated PRP at higher levels can be taken to reveal that the agent falls short of full-fledged ideal *responsibility* – rather than rationality. I have already suggested that, in its strictest sense, full-fledged ideally responsible agents are best represented by a model in which $R^2 = Id. = R^i$, for any $i > 2$. Clearly, Iterated PRP is fulfilled for $i > 2$.

Of course, R^2 may be symmetric (and hence an equivalence) without being the identity relation. In any case, it should be noted that, assuming $\subset\mathcal{M}$ satisfies clauses (1) to (8) from section 6, having $P^{i+1}{}_w(\underline{\mathbf{P}^{\mathbf{i}}(\varphi \mid \mathbf{P}^{\mathbf{i-1}}(\varphi|\ldots)\ldots.) = r}) = 1$ (and hence a symmetric R^i, for $i > 1$) is a *necessary* condition for the truth of $P^{i+1}{}_w(\varphi \mid \underline{\mathbf{P}^{\mathbf{i}}(\varphi \mid \mathbf{P}^{\mathbf{i-1}}(\varphi|\ldots\ldots.=s) = r})) = r$ (where *s* and *r* may not coincide) *but not a sufficient condition*. In addition, we should demand that higher-order *R*s do not "lose" further ordered pairs once R^{i-1} becomes an equivalence relation. More precisely, we can prove that:

(Necessary and sufficient condition for Iterated PRP)

$R^i = R^{i-1} \in \subset\mathcal{M}$ is an equivalence relation iff for all $w \in W$ and any $\underline{\varphi} \in L^0$, if $P^{i+1}{}_w(- \mid -)$ exists, then $P^{i+1}{}_w(\underline{\varphi} \mid \underline{\mathbf{P}^{\mathbf{i}}(\varphi \mid \mathbf{P}^{\mathbf{i-1}}(\varphi|\ldots)\ldots.) = r}) = r$

(See the Appendix).

This result tells us that Iterated PRP becomes true once a symmetric R^+ stabilizes. Hence, if satisfying Iterated PRP is taken to be important on independent grounds,[26] we have good reasons to ask for a stable R^+ as soon as possible. This can be seen as a strong motivation to adopt $R^2 = Id.$ for the representation of full-fledged ideally responsible agents, as suggested in section 6.

[26] For example, it has been contended that violations of (certain instances of) PRP make agents susceptible to Diachronic Dutch Books (DDB) (*cf.* for instance van Fraassen (1984), (1989), (1995)). It should be noted, however, that DDB arguments are even more controversial than PRP itself, so the rhetorical move from DDB to PRP needs to resort to additional considerations to be effective.

8. Other Approaches

There are a number of recent papers that also attempt to vindicate introspective principles. In particular, there are clear links between the model I have just presented and the formalism proposed by Paul Egré and Jérôme Dokic in (2008), and by Egré in (2009). Their main motivation is to deactivate Williamson's soritic argument against luminosity; in order to achieve their goal they distinguish between *perceptual* and *reflective* knowledge, each of which gets captured by a different operator. Their model then validates

(KK') $K_\pi\varphi \rightarrow KK_\pi\varphi$

(where "$K_\pi\varphi$" stands for "the agent has perceptual knowledge of φ". Thus, transparency failures at the perceptual level do not generalize to the reflective level.[27]

There are also some obvious differences between the two approaches. Egré/ Dokic do not offer a probabilistic framework, and they are concerned with reflections on perceptual knowledge, exclusively – hence knowledge operators beyond the second level are not allowed. Finally, the ***KK'*** principle is not a consequence of independent decisions on the formal structure of the system. In any case, the model for quasi-transparency presented here can be seen as a refinement of the system proposed by Egré and Dokic.

In McHugh (2010) we find another recent work whose analysis can be illuminating in connection with some of the ideas developed here, even though his proposal does not proceed at a formal level. McHugh (2010) argues that there is remarkable conceptual closeness between knowing that one believes and knowing that one knows, from which he takes the ***KK*** principle to follow; its truth – McHugh contends – does not hinge upon an internalist theory of justification, but just on the peculiar structure of self-knowledge. As opposed to McHugh's proposal, however, our framework can well tolerate a certain amount of transitivity failure, but without generalizing it to higher-order levels (which are indeed related to self-knowledge phenomena).

[27] Egré and Dokic show that Williamson's soritic argument is blocked once operators K and K_π are suitably distinguished from each other.

9. Conclusions

In the first part of this paper I suggested that transparency is not a demand of rationality, but of ideal responsibility, and hence that ideally responsible agents verify transparency principles. I also argued that the appropriate reflective stance required by ideal responsibility need not collapse with a justification stance; hence the satisfaction of reflective principles is not meant to be tied to an internalist epistemology.

In any case, the central argument of the paper in favor of transparency was addressed along sections 3 to 8, and proceeded indirectly through the development of a formal system. The formal framework presented here might have some interest on its own, I think, independently of its being conducive to the validation of introspective truths. Among other things, insofar as iterated knowledge operators belong to increasingly richer languages, the framework as a whole vindicates the idea that higher-order knowledge is crucially different from first-order knowledge. This is in agreement with a number of independent intuitions. In the first place, the attitude we adopt towards the fact that agents are typically more or less ignorant of the world (i.e., at the first level of knowledge) is normally very different from the attitude we adopt towards their ignorance at higher levels. If the considerations discussed in the first part of the paper are on the right track, we tend to assume that ideally responsible agents are aware of their own epistemic states – we *demand* them to be so aware – whereas we neither assume nor demand that agents be empirically omniscient. Second, resorting to different operators is in agreement with the intuition that second-order knowledge does not make room for "margin of error" principles (as Egré and Dokic (2008) were right to point out), and, more generally, it is in agreement with the idea that second-order knowledge is basically concerned with the possible "ratifiability" of first-order states.

The core of the formal argument relied on an attempt to make probabilistic and knowledge claims fit with each other smoothly. I showed that, once we understand that higher-order evidential probabilities require conditionalization over higher-order bodies of evidence, a coherent epistemic framework will lead us to validate several introspective principles; I have dubbed them *quasi-transparency* principles. It should be emphasized that quasi-transparency principles were not just assumed to hold, but they have been obtained as a result of implementing a number of natural constraints on the

structure of the system. Thus, formally speaking they behave quite differently from presuppositions of consistency or deductive closure.

These results vindicate the intuition put forward in section 2: even though second-order knowledge is not a demand of rationality, it is nonetheless an important desideratum of ideal epistemic subjects. It is an ideal we seek to fulfill to conceive of ourselves, not merely as rational creatures, but as full-fledged *agents*.[28]

[28] **Acknowledgments:** Thanks are due to Horacio Arló-Costa, Paul Egré, Ignacio Viglizzo and Timothy Williamson. Previous versions of this paper have been presented at talks and workshops held at the IIF (UNAM, Mexico), Universidad de Bahía Blanca (Bahía Blanca, Argentina), Universidad Torcuato Di Tella (Argentina), GAF (Universidad de Buenos Aires, Argentina), and UFESP (Sao Paulo, Brazil). I am very grateful to the participants for their comments and suggestions.

References

Bilgrami, A. (1999). "Why is self-knowledge different from other kinds of knowledge?" in L. E. Hahn (ed.), *The Philosophy of Donald Davidson*. Chicago: Open Court, pp. 211-224.

Christensen, D. (2004). *Putting Logic in its Place. Formal Constraints on Rational Belief.* Oxford: Clarendon Press.

Dokic, J. and Egré, P. (2009). "Margin for error and the transparency of knowledge." *Synthese* 166: 1-20.

Egré, P. (2008). "Reliability, margin for error, and self-knowledge." In Hendricks and Pritchard (2008) (eds.), pp. 215-250.

Engel, P. (mn.) "Epistemic responsibility without epistemic agency."

Foley, R. (2001). *Intellectual Trust in Oneself and Others.* Cambridge: Cambridge University Press.

Gaifman, H. (1986). "A theory of higher-order probabilities." In Brian Skyrms and William L. Harper (eds.), *Causation, Chance, and Credence.* Dordrecht: Kluwer Academic Publishers .

Hendricks, V. and Pritchard, D. (eds.) (2008). *New Waves in Epistemology*. Hampshire - New York: Palgrave Macmillan.

Hieronymi , P. (2005) "The wrong kind of reason." *Journal of Philosophy* 9: 437-57.

Hintikka, J. (1962), *Knowledge and Belief*. Ithaca: Cornell University Press.

Leitgeb, H. (2002). "Critical study of *Knowledge and its Limits." Grazer Philosophische Studien* 65: 195-205.

Levi, I. (1997). *The Covenant of Reason*. Cambridge: Cambridge University Press.

Nozick, R. (1981). *Philosophical Investigations*. Cambridge, Mass.: Harvard University Press.

McHugh, C. (2010). "Self-knowledge and the KK Principle." *Synthese* 173: 231-257.

Owens, D. (2000). *Reason without Freedom: The Problem of Epistemic Normativity*. London: Routledge.

Rescher, N. (2005). *Epistemic Logic. A Survey of the Logic of Knowledge*. Pittsburgh: University of Pittsburgh Press.

Samet, D. (1997). "On the triviality of high-order probabilistic beliefs".

Skyrms, B. (1980). “Higher order degrees of belief”. In D. H. Mellor (ed.) *Prospects for Pragmatism. Essays in honor of F. P. Ramsey*. Cambridge: Cambridge University Press.

Sosa, E. (2007). *A Virtue Epistemology: Apt Beliefs and Reflective Knowledge*, Vol. I. Oxford: Clarendon Press.

van Fraassen, B. (1984). “Belief and the Will.” *Journal of Philosophy* 81: 235-256.

van Fraassen, B. (1989). *Laws and Symmetry*. Oxford: Oxford University Press.

van Fraassen, B. (1995). “Belief and the problem of Ulysses and the Sirens", *Philosophical Studies* 77: 7-37.

Williamson, T. (ms). “Improbable knowing.”

Williamson, T. (1995). “Is knowing a state of mind?”. *Mind* 104: 533-565.

Williamson, T. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.

Appendix

Let $\supset\mathcal{M} = \langle W, R^1, \ldots R^n \ldots, P_{prior}, v\rangle$ satisfy requirements (1) to (8) from section 6. Then:

1. (K) For $\underline{K^1(\varphi \rightarrow \psi)}$, $\underline{K^1\varphi}$, and $\underline{K^1\psi} \in L^1$: $\underline{K^1(\varphi \rightarrow \psi) \rightarrow (K^1\varphi \rightarrow K^1\psi)}$ is valid in $\supset\mathcal{M}$.

Proof: Trivial from the definition of true-in-a-world for sentences with K^i operators, and the fact that $\supset\mathcal{M}$ validates *Modus Ponens*.

[Suppose $\vDash_w \underline{K^1(\varphi \rightarrow \psi)}$, $\vDash_w \underline{K^1\varphi}$, and not-$\vDash_w \underline{K^1\psi}$ (for *reductio*). Then there is some $x \in W$ such that wR^1x, $\vDash_x \underline{\varphi \rightarrow \psi}$, $\vDash_x \underline{\varphi}$, and not-$\vDash_x \underline{\psi}$, which is impossible. Thus $\vDash_w \underline{K^1(\varphi \rightarrow \psi) \rightarrow (K^1\varphi \rightarrow K^1\psi)}$, for any $w \in W$.] ■

Notice that $\underline{K^i(\varphi \rightarrow \psi) \rightarrow (K^i\varphi \rightarrow K^1\psi)}$ holds vacuously in $\supset\mathcal{M}$ for any $i > 1$, given that, for $i > 1$, $\underline{K^i(\varphi \rightarrow \psi)}$ is not a wff of $L^0, \ldots L^n \ldots$. Thus, well-discussed instances of **K** in other normal systems, such as (*)$(\underline{K^2(Kp \rightarrow p) \rightarrow (K^2Kp \rightarrow K^2p)}$, are not instances of **K** in this model, insofar as neither (*)$\underline{K^2(Kp \rightarrow p)}$ nor (*)$\underline{K^2p}$ are wff in $\supset\mathcal{M}$.

2. (KK$^+$) For any $\underline{K^2\varphi} \in L^2$: $\underline{K^2\varphi \rightarrow K^3K^2\varphi}$ is a valid formula in $\supset\mathcal{M}$.

Proof: Trivial from the fact that R^2 is transitive and the Rs are nested.

[Assume $\vDash_w \underline{K^2\varphi}$, for $w \in W$. Suppose (for *reductio*) that not-$\vDash_w \underline{K^3K^2\varphi}$. Hence there is $x \in W$ such that wR^3x and not-$\vDash_x \underline{K^2\varphi}$, which means that there is some y such that xR^2y and not-$\vDash_y \underline{\varphi}$. As R^2 is transitive and higher Rs never add pairs, we also have wR^2x and wR^2y, and hence not-$\vDash_w \underline{K^2\varphi}$, contrary to our assumption. Hence $\vDash_w \underline{K^2\varphi \rightarrow K^3K^2\varphi}$, for any $w \in W$.]■

Corollary I: For any $\underline{K^i\varphi} \in L^i$ and $i > 1$: $\underline{K^i\varphi \rightarrow K^{i+1}K^i\varphi}$ is a valid formula in $\supset\mathcal{M}$.

[Trivial from the fact that R^i is transitive, for any $i > 1$, and the fact that the Rs are nested].

Corollary II: If R^1 is transitive, $\underline{K^i\varphi \rightarrow K^{i+1}K^i\varphi}$ is valid for any $i \geq 1$.

3. (T) For any $i \geq 1$ and any $\underline{K^i\varphi} \in L^i$: $\underline{K^i\varphi \rightarrow \varphi}$ is valid in $\subset\mathcal{M}$.

Proof: Trivial from the fact that R^i is reflexive, for any $i \geq 1$.

Corollary: For all $i > 1$, and any $\underline{\varphi} \in L^0$: $\underline{K^iK^{i-1}\ldots K\varphi \leftrightarrow K^2K\varphi}$ is valid in $\subset\mathcal{M}$, even though $R^2\ldots$, $R^i\ldots$ may well be distinct. [Straightforward from **(KK$^+$)** and **(T)**]

4. $\underline{K\varphi \rightarrow K^2K\varphi}$ is not generally valid in $\subset\mathcal{M}$.

Proof: We can build a straightforward counterexample. Let R consist of pairs of worlds (w, w), (w, x), (x, x), (x, y), and (y, y), but not (w, y); let R^2 be obtained from R by eliminating (x, y). Assume further $\vDash_w\underline{\varphi}$, $\vDash_x\underline{\varphi}$ and not-$\vDash_y\underline{\varphi}$. Hence we have both $\vDash_w\underline{K\varphi}$ and not-$\vDash_x\underline{K\varphi}$. As (w, x) is still in R^2, not-$\vDash_w\underline{K^2K\varphi}$. ■

5. (KK^2) Let $R^2 \in \subset\mathcal{M}$ satisfy the following property:

(+) $\forall w\forall x \in W\,(wR^2x \rightarrow x \in [KR(w)])$

Then, for any $\underline{K\varphi} \in L^1$, $\underline{K\varphi \rightarrow K^2K\varphi}$ is a valid formula in $\subset\mathcal{M}$.

Proof: Assume the property holds, and suppose both $\vDash_w\underline{K\varphi}$ and not-$\vDash_w\underline{K^2K\varphi}$, for $w \in W$ (for *reductio*). Then there exists some y such that wR^2y and not-$\vDash_y\underline{K\varphi}$. As (+) applies, $y \in [KR(w)]$, hence $\vDash_yR(w)$. Given that $\underline{K\varphi}$ is true in w, we have $[R(w)] \subseteq [K\varphi]$, and therefore $\vDash_y\underline{K\varphi}$, which is impossible. Hence $\vDash_w\underline{K\varphi \rightarrow K^2K\varphi}$, for any $w \in W$. ■

Corollary I: Let $R^2 \in \subset\mathcal{M}$ satisfy property (+). Then for all $i > 0$, and any $\underline{\varphi} \in L^0$: $\underline{K^iK^{i-1}\ldots K\varphi \leftrightarrow K\varphi}$ is valid in $\subset\mathcal{M}$.

[Straightforward from **5.** and the **Corollary** of **3**]

Corollary II: Let $R^2 \in \subset\mathcal{M}$ be the identity relation (*Id.*). For any $\underline{K\varphi} \in L^1$, $\underline{K\varphi \rightarrow K^2K\varphi}$ is a valid formula in $\subset\mathcal{M}$.

[Straightforward from **5** and the fact that, for any $w \in W$: $w \in [KR(w)]$, hence *Id.* satisfies property (+)]

6. (KK$^{\Diamond}$) Let $R^2 \in \mathcal{C}\mathcal{M}$ satisfy the following property:

(++) $\exists w \forall x \in W (wR^2x \rightarrow x \in [KR(w)])$

Then there is some $w \in W$ such that $\vDash_w \underline{K\varphi} \rightarrow \underline{K^2K\varphi}$, for any $\underline{K\varphi} \in L^1$.

Proof: Straightforward from **5**. ■

7. For any $w \in W$, any $\varphi \in L^0$, and any $r \in [0, 1]$, if R^1 is an equivalence relation, $P^2{}_w(\underline{\mathbf{P}^1(\varphi)=r})$ is either 1 or 0.

Proof: If R^1 is an equivalence relation over W, then for any $x \in [R(w)]$: $[R(x)] = [R(w)] = [KR(w)] = [KR(x)]$, hence $P^1{}_w(\underline{\varphi}) = P_{prior}(\underline{\varphi} \mid \underline{R(w)}) = P_{prior}(\underline{\varphi} \mid \underline{R(x)}) = P^1{}_x(\underline{\varphi})$, and, moreover, $P^2{}_w(\underline{\mathbf{P}^1(\varphi)=r}) = P_{\text{prior}}(\underline{\mathbf{P}^1(\varphi)=r} \mid \underline{KR(w)}) = P_{\text{prior}}(\underline{\mathbf{P}^1(\varphi)=r} \mid \underline{KR(x)}) = P^2{}_x(\underline{\mathbf{P}^1(\varphi)=r})$. Now, $P_{\text{prior}}(\underline{\mathbf{P}^1(\varphi)=r} \mid \underline{KR(w)}) = \#\{y \in W: \vDash_y \underline{\mathbf{P}^1(\varphi)=r} \ \& \ \vDash_y \underline{KR(w)}\} / \#\{y \in W: \vDash_y \underline{KR(w)}\}$. But, as we have seen, all y in $[KR(w)]$ coincide in the probability they give to $\underline{\varphi}$; in particular, either all of them give $\underline{\varphi}$ probability r, or none does. Hence, either all y in $[KR(w)]$ make $\underline{\mathbf{P}^1(\varphi)=r}$ true, or none does. Thus $P^2{}_w(\underline{\mathbf{P}^1(\varphi)=r})$ is either 1 or 0. ■

Corollary I: For $i > 1$, if R^i is symmetric, then $P^{i+1}{}_w(\underline{\mathbf{P}^i(\ldots\mathbf{P}^1(\varphi)\ldots)=r})$, if it exists, is either 1 or 0, for any $w \in W$, any $\varphi \in L^0$, and any $r \in [0, 1]$.

[Straightforward from **7.** and the fact that, for any $i > 1$, R^i is also reflexive and transitive, hence R^i is an equivalence relation and partitions W).

Corollary II: Let R^i be the first equivalence relation in $R^1 \ldots R^n \ldots$Then for $j \geq 1$, any $w \in W$, and any $\underline{\varphi}$ in the P-fragment of L^{i+j-1}: $P^{i+j}{}_w(\underline{\varphi})$, if it exists, is always 1 or 0.

8. For any $\varphi \in L^0$, if $P^1{}_w(\underline{\varphi}) = r = 1$ or 0, then $P^2{}_w(\underline{\mathbf{P}^1(\varphi)=r}) = 1$ (regardless of how R^1 is).

Proof: Suppose $P^1_w(\underline{\varphi}) = 1$. Then we have $\vDash_x \underline{\varphi}$, for all $x \in [R(w)]$. Moreover, for every $y \in [KR(w)]$ and all z such that yRz, $\vDash_z \underline{R(w)}$, i.e., $z \in [R(w)]$. Hence $\vDash_z \underline{\varphi}$, and $\vDash_y \underline{\mathbf{P}^1(\varphi)=1}$. As this is the case for every y in $[KR(w)]$, $P^2_w(\underline{\mathbf{P}^1(\varphi)=1}) = 1$.

Symmetrically, suppose now $P^1_w(\underline{\varphi}) = 0$. Then for all $x \in [R(w)]$, not-$\vDash_x \underline{\varphi}$. Again, for every $y \in [KR(w)]$ and all z such that yRz, $\vDash_z \underline{R(w)}$, hence not-$\vDash_z \underline{\varphi}$, and $\vDash_y \underline{\mathbf{P}^1(\varphi)=0}$. As this holds for every y in $[KR(w)]$, $P^2_w(\underline{\mathbf{P}^1(\varphi)=0}) = 1$. ■

Corollary: If $P^i_w(\underline{\varphi}) = r = 1$ or 0, for any $\underline{\varphi}$ in the P-fragment of L^{i-1}, then for any $j \geq 1$: if $P^{i+j}_w(\text{-})$ exists, $P^{i+n}_w(\underline{\mathbf{P}^i(\varphi)=r}) = 1$.

9. Suppose $P^2_w(\underline{\mathbf{P}(\varphi)=r}) = s$. If $0 \neq r \neq 1$ and R^1 is not transitive, then s need not be either 1 or 0.

Proof: For a counterexample, suppose $W = \{w, x, y, z\}$. Let R^1 be reflexive and symmetric, but not transitive; let (x, w), (x, y) and (y, z) be in R^1, but not (x, z); let also $[\varphi] = \{w\}$. Then $P^1_x(\underline{\varphi}) = 1/3$, and $P^2_x(\underline{\mathbf{P}(\varphi)=1/3}) = 1/2$. ■

10. (Necessary and sufficient conditions for Iterated PRP)

$R^i = R^{i-1} \in \mathcal{M}$ is an equivalence relation iff for all $w \in W$ and any $\underline{\varphi} \in L^0$, if $P^{i+1}_w(\text{- | -})$ exists, then $P^{i+1}_w(\underline{\varphi} \mid \underline{\mathbf{P}^i(\varphi \mid \mathbf{P}^{i-1}(\varphi \mid \ldots)\ldots.) = r}) = r$

Proof: If R^{i-1} is an equivalence relation and $R^i = R^{i-1}$, Williamson's results apply and Iterated PRP holds. Williamson's results also tell us that the principle does not hold when R^{i-1} is not an equivalence relation. Thus we only need to show that, if $R^i \neq R^{i-1}$, then there is some world in which, if Iterated PRP is well defined, it is false, even if both R^i and R^{i-1} are equivalence relations. If R^{i-1} is an equivalence relation but $R^i \neq R^{i-1}$, there are some w, x such that $x \in [K^{i-1}\ldots R(w)]$ but $x \notin [K^i\ldots R(w)]$. Assume that $P^{i+1}_w(\text{- | -})$ exists, and hence that $[K^i\ldots R(w)]$ is not empty. Let $[\varphi] = \{x\}$. Then $P^{i+1}_w(\underline{\varphi} \mid \underline{\mathbf{P}^i(\varphi \mid \ldots.) = r}) = s$, where $r = 1 / \#[K^{i-1}\ldots R(w)]$, and $s = 0$ (because there is no world in $[K^1\ldots R(w)]$ in which $\underline{\varphi}$ is true).■